A bioinformatics workflow for sequence classification beyond species-level resolution

We present a bioinformatics workflow designed to increase the taxonomic resolution of sequence classification beyond species-level. The workflow comprises three components: RepGenR, for downloading whole genome sequences and filtering non-informative or redundant representatives; FlexMetR, for integrating high-resolution information such as subspecies, antibiotic resistance and virulence factors; and FlexTaxD, for merging and refining taxonomies for use with classification tools like Kraken2.

We also apply this to *Francisella tularensis* and show results on classification an environmental sample.

Authors

Caroline Öhrman, Jacob Lewerentz, David Sundell, Edvin Karlsson, Kerstin Myrtennäs and Andreas Sjödin.

FOI, Swedish Defence Research Agency

1 Select pathogen(s) of interest

The highly virulent *Francisella tularensis* subsp. *tularensis* (Type A) is genetically similar to the less virulent subsp. *holarctica* (Type B), a distinction often missed by standard reference databases, including the Genome Taxonomy Database (GTDB).

This bioinformatics workflow (Figure 1), divided into three components integrates subspecies and genotype information of *F. tularensis* into a database suitable for DNA sequence read classification. This method is generic and can be applied at any taxonomic level within the GTDB.

2 Systematic replacement of nodes

The GTDB representative genomes were used as a base, with genera within *Francisellaceae* being separately replaced. Species and subspecies were supplemented with a dereplicated set of *F. tularensis* genomes at 99.95% ANI. The systematic process of iteratively replacing nodes recursively using the workflow is illustrated in Figure 2.

The enhanced *Francisellaceae* database is visualized in Figure 3. The metadata added to nodes and branches are genus, species, subspecies and clade information – for example, "Francisella tularensis tularensis A.I.1" representing the highly virulent type of Type A tularemia.

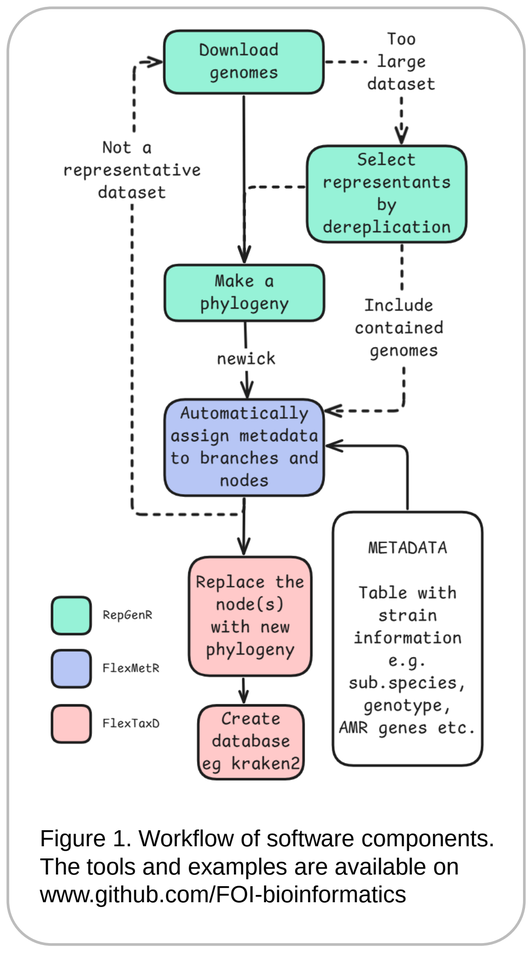

Figure 1. Workflow of software components. The tools and examples are available on www.github.com/FOI-bioinformatics

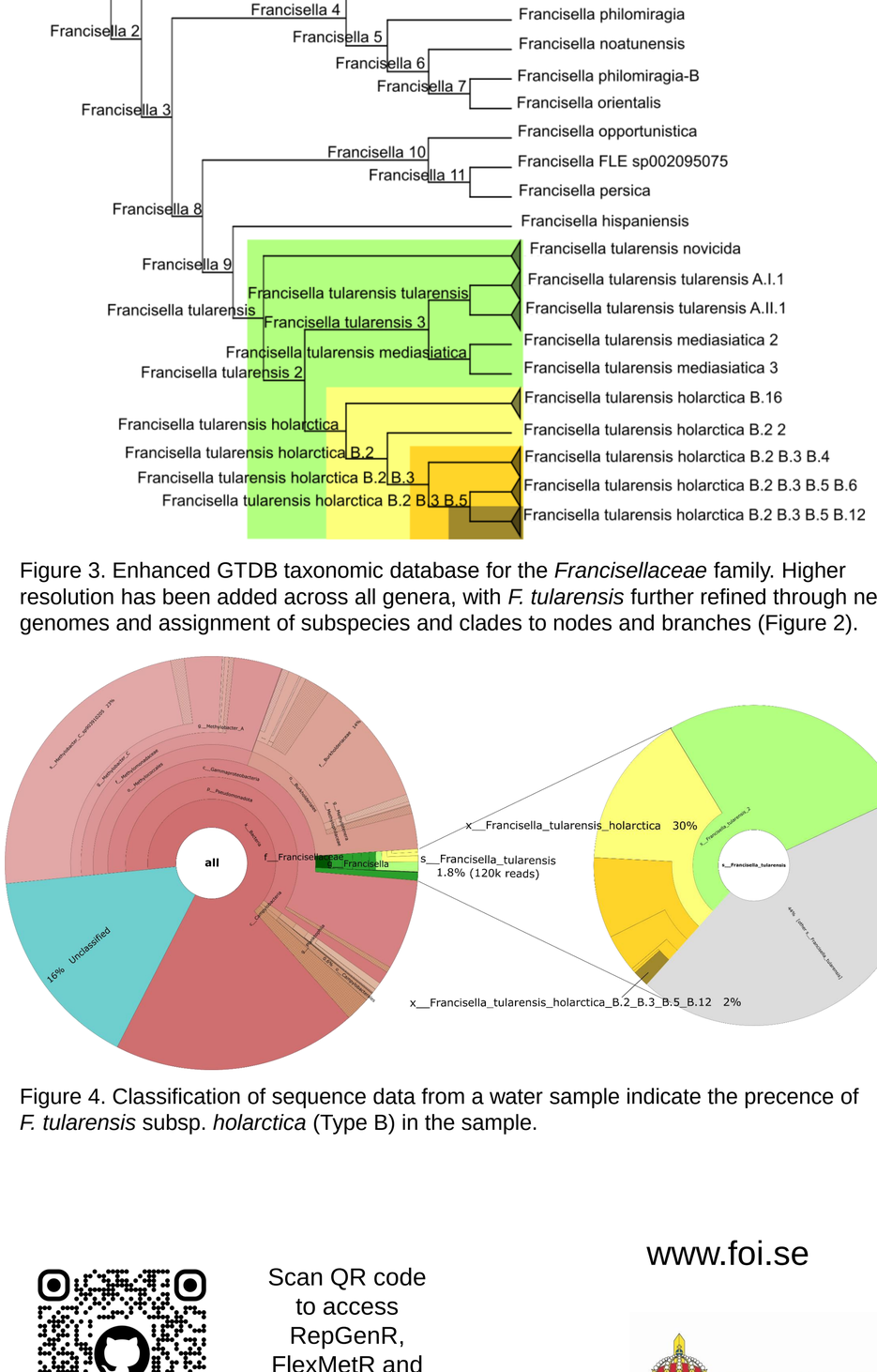

Figure 3. Enhanced GTDB taxonomic database for the *Francisellaceae* family. Higher resolution has been added across all genera, with *F. tularensis* further refined through new genomes and assignment of subspecies and clades to nodes and branches (Figure 2).

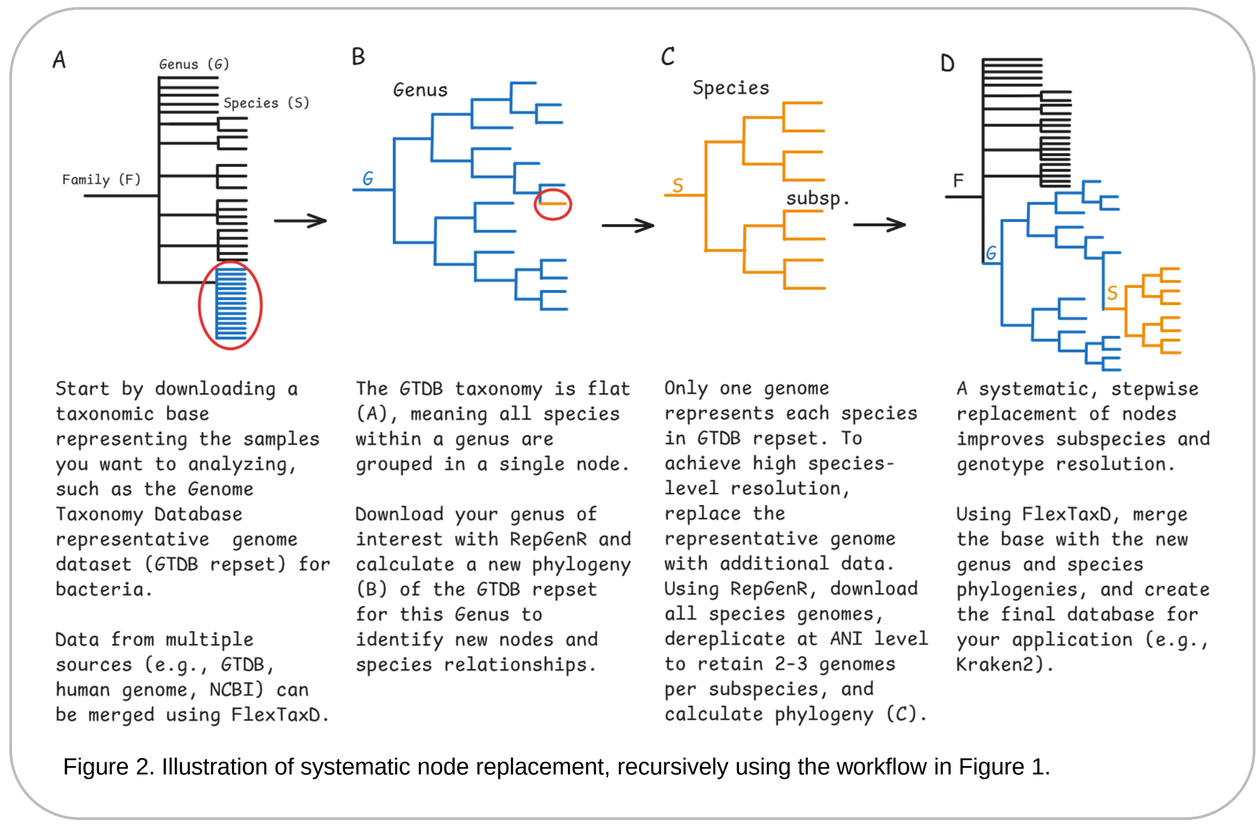

Figure 2. Illustration of systematic node replacement, recursively using the workflow in Figure 1.

3 Classify sequence data

Sequence reads from a drinking water sample connected to tularemia cases in northern Sweden (2024) were classified using kraken2 with the enhanced database.

Of the total reads, 2% were classified as *F. tularensis* at the species level using the GTDB database alone. Further analysis of these reads revealed strong support (30%) for *F. tularensis* subsp. *holarctica* (Type B), specifically indicating clade B.12 – a genotype commonly found in this region. This genotype identification was subsequently confirmed using additional methods.

Figure 4. Classification of sequence data from a water sample indicate the precence of *F. tularensis* subsp. *holarctica* (Type B) in the sample.

Conclusion

This bioinformatics workflow creates databases with beyond-species taxonomic resolution, applicable on-demand to any Genome Taxonomy Database (GTDB) taxon while utilizing the latest genomic resources. Testing with an enhanced *Francisella tularensis* database indicates its applicability for subspecies classification in metagenomic sequence data.

Scan QR code to access RepGenR, FlexMetR and FlexTaxD on github.

www.foi.se

FOI